Vol.2021-NL-250 No.3
2021/9/28

抽出型複数文書要約における文順序を考慮した評価

藤田 正悟[1]　上垣外 英剛[1]　船越 孝太郎[1]　奥村 学[1]

概要：抽出型要約は元の文書において重要度が高い文を抽出し要約として再構成する手法であり広く使われている．その一方，この方法では複数文書を横断して重要文を抽出し要約を構成する際に，抽出された文の順序が適切ではない場合がある．解決策として既存の文並び替えモデルを使って文並び替えを行うことが考えられるが，抽出型要約に適した文並び替えの教師データが存在せず，尚且つ抽出型要約において並び替えを考慮した評価尺度が存在しないという問題がある．そこで我々は抽出型要約に適した文並び替えの教師データの作成手法と抽出型要約を並び替える場合の評価指標を提案する．いくつかのベースラインと比較した結果，我々の評価指標は特に一貫性において人手評価と高い相関を示した．

1. はじめに

抽出型文書要約は，元の文書において重要度が高い文を抽出し要約として再構成する．近年，このタスクに深層ニューラルネットを用いた研究が盛んに行われ，高い精度が報告されている [6], [7], [14], [17], [21], [26]．その中でも複数文書を横断して文を抽出して要約文書を構成する複数文書要約は，従来の単文書要約とは異なる傾向が見られる [8]．単文書での抽出型要約を行う手法でも高い精度で複数文書要約を行うことができることが報告されている [26]．単文書要約手法は，要約文を抽出した後にそれらを元の文書の順番に並び替える後処理を行うのが一般的である．そのため，複数文書をつなげて擬似的に単文書として入力する場合には抽出した要約文書の文の並びが適切でないことがある．要約文書の文の並びが不適切になる例として，複数文書要約データセット Multi-News[8] に含まれる抽出型要約の oracle を元の文書の順番で並べた図 1 が挙げられる．1 文目の内容を補足するオレンジ色の文が 5 文目となってしまっており，読み手はどのような事件についての裁判なのかがわからないまま 5 文目まで読むことになり，読みづらい文章となってしまっている．

このような問題を解決するためには，抽出型要約モデルが文を抽出した後にその文書を適切に並び替える手法を導入することが考えられる．しかし，文並び替え手法の教師データとして必要な正しい要約文書の順序がわからない．そのため，既存の文並び替え手法を抽出した要約文書を並び替えるタスクにそのまま適用することができない．そこで，我々は人手で生成した要約文書を基に並び替えの教師データを作成する方法と，抽出型要約の文並び替えを評価する指標を提案する．この文並び替え評価指標は既存の評価指標と比べて人手評価との強い相関を示した．

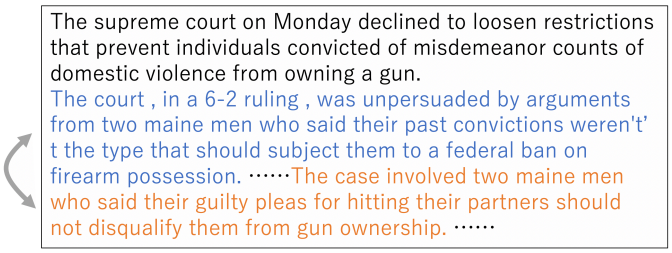

図 1　抽出型要約による読みづらい要約結果

2. 関連研究

抽出型文書要約は元の文書から関連する文章を抽出し要約として再構成する手法であり，言語モデルを利用した要約モデルが多く提案されている [6], [7], [14], [17], [21], [26]．さらに，現在は単文書の要約だけでなく複数文書の要約も盛んに行われている [8], [10]．複数文書の要約データセットとして Multi-News[8] が挙げられる．Multi-News は複数の記事を結合した文書と人手で生成した要約文書からなるデータセットである．また，文書の一貫性を保つ手法は文章の読みやさを確保するために重要であり，多くの研究が行われている [9], [15], [16], [22]．その中でも並び替えタスクはよく取り組まれている分野であり [2], [3], [11]，近年はニューラルモデルを用いた文並び替えモデルが提案されている [4], [5], [12], [19], [23], [24].

抽出型要約モデルの出力は ROUGE[13] を用いて評価するのが一般的である．また，文生成タスクの評価指標として BERTScore[25] が使われている．一方で，文並び替えを

[1] 東京工業大学
Tokyo Institute of Technology

評価する場合，ROUGE はほとんど文の並び替えの影響を受けず，BERTScore は単語のアライメントを評価しているため文の並びを評価しているとは言えず，抽出型要約において並び替えを考慮した評価尺度が存在しない．

また，文単位の並び替えの評価には，人手で作成した並び順との一致度合いを評価指標として用いるのが一般的である [2], [3], [11]．しかし，抽出型要約モデルの出力を人手で並び替えるのは高いコストがかかる．そこで，我々は人手で作られた要約文書と抽出型要約モデルの出力の対応を考えることで抽出型要約における文順序を評価する手法を提案する．

3. 提案手法

3.1 抽出型複数文書要約のタスク定義

抽出型複数文書要約は 2 段階のタスクである．(1) 入力として文書が与えられ，抽出型要約モデルが要約文書を抽出する．(2) 次に，文並び替えモデルで (1) の出力を読みやすい順番に並び替える．

本論文の実験では，3.2 節で説明するアルゴリズムを用いて作ったデータを教師データとして文並び替えモデルを学習する．また，並び替えられた要約文書は 3.3 節で説明する評価指標を用いて評価する．

3.2 文並び替えモデルの正例の作成

Algorithm 1 文並び替えモデルの正例の作成

Require: all_comb $\Leftarrow \{\}$, matching $\Leftarrow \{\}$
  $P \Leftarrow$ 複数文書を横断して抽出された要約構成文の集合,
  $G \Leftarrow$ 人手で生成した要約文書の構成文集合
  for each p in P **do**
    for each g in G **do**
      all_comb $\Leftarrow$ all_comb $\cup$ (BERTScore$(p, g), p, g)$
    end for
  end for
  all_comb を BERTScore の降順に並び替える
  for each comb in all_comb **do**
    if comb.p を含むペアを既に選んでいる **then**
      continue
    end if
    if comb.g を含むペアを既に選んでいる **then**
      if G の中でまだ $matching$ に含まれていない文がある **then**
        continue
      end if
    end if
    matching $\Leftarrow$ matching $\cup$ (comb.g,comb.p)
  end for
  matching に従い，G 内の順番に従って P を並び替える

アルゴリズム 1 で並び替えた文書を文並び替えモデルの教師データとして利用する．このアルゴリズムは人手で生成した要約文書と抽出した要約文書の各文で BERTScore を計算し，BERTScore が高い文のペアから順番に採用する．全ての文を含むペアを選び終えると，それをアライメントとして，抽出した要約文書を並び替える．この操作により抽出した要約文書を人手で生成した要約文書の並びに近くなるように並び替えることができる．

3.3 評価指標

我々は，抽出型複数文書要約を評価する手法として Ordered-BERTScore-v1 と Ordered-BERTScore-v2 の 2 つを提案する．これらの指標はパラメータ n によって制御することができる．

まず，評価指標 Ordered BERTScore-v1-n=1 について説明する．Ordered BERTScore-v1-n=1 は人手で生成された要約文書 (gold) と出力文書 (prediction) の 1 文対 1 文のアライメントに基づいてスコアを算出する．まず，gold と prediction のそれぞれの文での BERTScore を計算する．BERTScore は入力した 2 文の単語アライメントを元に推定した類似度である．BERTScore の値域は $[-1, 1]$ であり，類似度が高いほど値も大きくなる．BERTScore-v1-n=1 は漸化式

$$S[i][j] = max \begin{cases} S[i][j-1] \\ S[i-1][j] \\ S[i-1][j-1] + C_{i,j} \end{cases} \quad (1)$$

に従って，gold の各文と prediction の各文をアライメントをとった場合の BERTScore の合計値を計算する．S は動的計画法のテーブルであり，$C_{i,j}$ はテーブルの縦軸の i 番目の文と横軸の j 番目の文の BERTScore である．gold を縦軸として計算した場合のテーブルは図 2 の左側のようになる．

gold の文の数を $|g|$，prediction の文の数を $|p|$ とする．gold を縦軸として計算したときの $S[|g|][|p|]$ を S_{rec}，prediction を縦軸として計算したときの $S[|p|][|g|]$ を S_{prec} とすると，Ordered-BERTScore-v1-n=1 は，$\frac{S_{rec}}{|g|}$ と $\frac{S_{prec}}{|p|}$ の調和平均である．

次に，1 文対複数文のアライメントに基づいてスコアを算出する一般の場合の Ordered BERTScore-v1-n ($n > 1$) について説明する．Ordered BERTScore-v1 は漸化式

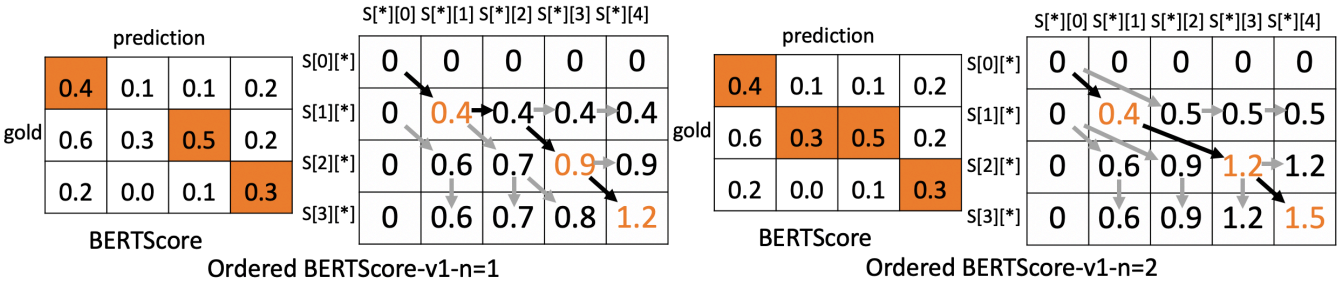

図 2 Ordered BERTScore-v1 の動的計画法テーブル．太字の矢印が採用された経路であり，オレンジの文字が値の更新があった場所である．BERTScore の表の濃いオレンジで囲まれた部分がアライメントである．

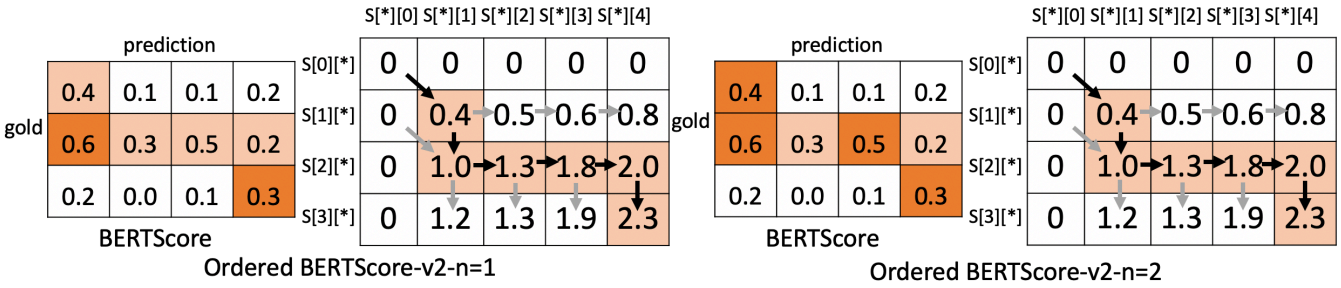

図 3 Ordered BERTScore-v2 の動的計画法テーブル．太字の矢印が採用された経路であり，薄いオレンジで囲まれた場所が選ぶペアの候補である．BERTScore の表の濃いオレンジで囲まれた部分がアライメントである．

$$S[i][j] = max \begin{cases} S[i][j-1] \\ S[i-1][j] \\ S[i-1][j-1] + C_{i,j} \\ S[i-1][j-2] + C_{i,j} + C_{i,j-1} \\ \vdots \\ S[i-1][j-n] + \sum_{k=0}^{k=n-1} C_{i,j-k} \end{cases} \quad (2)$$

に従って計算する．$n = 2$ の時に gold を縦軸として計算した場合のテーブルは図 2 の右側のようになる．Ordered BERTScore-v1 も $\frac{S_{rec}}{|g|}$ と $\frac{S_{prec}}{|p|}$ の調和平均である．n を大きくすることで，1 文が長く様々な情報が含まれる文がある場合や，類似する文が多い場合に適切に評価することができるようになる．

次に，複数文対複数文のアライメントに基づいてスコアを算出する Ordered BERT-v2-n について説明する．Ordered BERT-v2 は，漸化式

$$S[i][j] = max \begin{cases} S[i][j-1] + C_{i,j} \\ S[i-1][j] + C_{i,j} \\ S[i-1][j-1] + C_{i,j} \end{cases} \quad (3)$$

に従って S を計算する．次に，動的計画法の経路上のペアからスコアが大きい順に，各行と各列で n 個より多く重複しないように選ぶ．例えば図 3 の左側の例の場合は，$(2,1),(3,4)$ の順番で選び，図 3 の右側の例の場合は，$(2,1),(2,3),(1,1),(3,4)$ の順番で選ぶ．Ordered BERTScore-v2 は，選んだペアの BERTScore の和を sum_{v2} とすると，$\frac{sum_{v2}}{|p|+|g|-1}$ である．

Ordered BERTScore-v2 は Ordered BERTScore-v1 に比べて，文の対応関係が綺麗にとれない場合でも評価できるという利点がある．文の対応上関係が綺麗にとれないというのは，gold の 1 文目が A, B という情報を，gold の 2 文目が C, D という情報を持っており，prediction の 1 文目が B, C という情報を，prediction の 2 文目が D, E という情報を持っているというような場合である．この場合，BERTScore-v2-($n > 1$) であれば，gold の 1 文目と prediction の 1 文目，gold の 2 文目と prediction の 1 文目，gold の 2 文目と prediction の 2 文目のペアでアライメントを取って評価することができる．

4. 実験

実験は Multi-News[8] をデータセットとして，2 つの設定で行った．1 つ目は，抽出型要約の oracle を文並び替えモデルで並べ替える設定である．2 つ目は，抽出型要約モデルの出力をさらに文並び替えモデルで並び替える設定である．

評価指標のベースラインとして，BERTScore[25]，GPT-2[20] の perplexity，ROUGE[13]，coherence[1], [12], [18], [23] を利用した．coherence は文書を s，s の i 文目の文

相関係数	spearman	kendalltau
BERTScore	0.21	0.19
GPT-2 perplexity	-0.03	-0.03
ROUGE-1	0.06	0.05
ROUGE-2	0.09	0.08
ROUGE-l	0.09	0.08
coherence-λ=0.0	-0.02	-0.02
coherence-λ=0.5	0.00	0.00
coherence-λ=0.8	0.07	0.06
coherence-λ=1.0	0.11	0.10
Ordered BERTScore-v1-n=1	0.22	0.20
Ordered BERTScore-v1-n=2	0.19	0.17
Ordered BERTScore-v1-n=3	0.27	0.25
Ordered BERTScore-v1-n=4	0.23	0.20
Ordered BERTScore-v1-n=∞	0.23	0.21
Ordered BERTScore-v2-n=1	0.20	0.18
Ordered BERTScore-v2-n=2	0.21	0.19
Ordered BERTScore-v2-n=3	0.26	0.23
Ordered BERTScore-v2-n=4	**0.33**	**0.29**
Ordered BERTScore-v2-n=∞	0.25	0.22

表 1 文並び替えのみの評価

ベクトルを s_i，NE(·) を文書の固有表現の集合を返す関数だとすると，

$$\text{coherence}(s) = \frac{\sum_{i=1}^{N-1} \text{Sim}(s_i, s_{i+1})}{N-1} \tag{4}$$

$$\text{Sim}(s_i, s_{i+1}) = \lambda \cdot \text{NESim}(s_i, s_{i,j}) + (1-\lambda) \cdot \text{CosSim}(s_i, s_{i,j}) \tag{5}$$

$$\text{NESim}(s_i, s_{i+1}) = \frac{|\text{NE}(s_i)| \bigcap |\text{NE}(s_{i+1})|}{min(|\text{NE}(s_i)|, |\text{NE}(s_{i+1})|)} \tag{6}$$

$$\text{CosSim}(s_i, s_{i+1}) = \frac{s_i \cdot s_{i+1}}{||s_i||\ ||s_{i+1}||} \tag{7}$$

と表せる．λ は $[0, 1]$ の値域を持つパラメータである．

また，提案手法の Ordered BERTScore-v1-n=∞,Ordered BERTScore-v2-n=∞ は，それぞれ $n = max(|g|, |p|)$ とした設定である．

4.1 文並び替えのみの評価

抽出型要約の oracle を文並び替えモデルで並べ替えたものを評価した (表 1)．文並び替えモデルとして，Cui らの手法 [5] と Zhu らの手法 [27] を使った．oracle を 2 つのモデルで並び替えたもの，ランダムに並び替えたもの，の 3 種類の出力を提案法を含む 19 種類の評価指標で評価した．人手評価は amazon mechanical turk にて，50 件のデータを 10 人の worker で行った．文書の一貫性を評価基準として 5 段階でつけてもらった．

結果として，既存指標に対して提案指標の方が人手評価と高い相関があった．また，Ordered BERTScore-v2 は Ordered BERTScore-v1 よりも相関が高かった．文の対応関係が綺麗にとれない場合が多かったため，それをうまく評価できる Ordered BERTScore-v2 の方が相関が高かったのだと考えられる．また，文書内に類似文が複数含まれるケースが多く見られたため Ordered BERTScore-v2 の中でも Ordered BERTScore-v2-n=4 が最も人手評価と相関が高かったと考えられる．

4.2 抽出型要約後の文並び替えの評価

抽出型要約モデルとして，Zhong らの手法 [26] と Wang らの手法 [21] を使い，文並び替えモデルは 1 つ目と同様に 2 つのモデルを使った．それぞれ 2 種類のモデルを用いた 2×2 通りの出力を提案法を含む 19 種類の評価指標で評価した．この設定は，50 件の文書を informativeness, readability, consistency の 3 つで評価してもらった．(表 2). informativeness は人手で作られた要約文書の情報を過不足なく含んでいるか，readability は文法的に正しいか，consistency は内容に一貫性があり読みやすいかという基準で評価してもらった．全体の傾向として，抽出型要約モデルの出力をさらに並び替えているため必ずしも出力が読みやすいとは言えず人手評価の分散が大きかった．また，人手で作られた要約文と要約器が抽出した要約文の BERTScore が負になる場合が多く見られた．

informativeness は Ordered BERTScore-v2-n=1 が最も人手評価との相関が高かった．これは BERTScore-v2 は BERTScore が負の場合でも評価することができ，特に BERTScore-v2-n=1 は BERTScore が著しく低いペアの影響を受けずに評価できるためだと考えられる．readability は BERTScore が最も人手評価と相関が高かった．BERTScore が提案手法よりも人手評価との相関が高かったのは，BERTScore は文書全体の単語アライメントを考えられるためだと考えられる．consistency は 4.1 節の設定と異なり，Ordered BERTScore-v2-n=1 が最も人手評価との相関が高かった．抽出した要約文書をさらに並び替える設定であるため gold と prediction の差異が大きく，複数文のアライメントを考えるよりも 1 文対 1 文のアライメントを考える方が適していたのだと考えられる. また，BERTScore-v1 は負の場合の大小を評価することができないので，BERTScore-v2-n=1 が最も優れていたのだと考えられる．

5. 分析

アルゴリズム 1 を用いて図 1 の文書を並び替えた例を紹介する (図 4)．1 文目の情報を補足するオレンジ色の文がすぐ後に来ており，その後に青色の判決についての文が来ている．そのため元の並びよりも読みやすくなっている．

また，4.1 節の設定での一貫性の人手評価は表 5 のようになった．Cui らの手法 [5] による並び替えが 5 段階評価で 4 以上をつけた人の割合が最も多く，平均値も最も高かった．また，ランダムに並び替える手法は明らかに他の 2 つ

評価基準	informativeness		readability		consistency	
相関係数	spearman	kendalltau	spearman	kendalltau	spearman	kendlltau
BERTScore	**0.04**	0.03	**0.11**	**0.09**	0.14	0.12
GPT-2 perplexity	-0.10	-0.08	-0.08	-0.06	-0.27	-0.22
ROUGE-1	0.02	0.02	0.05	0.04	0.03	0.03
ROUGE-2	-0.06	-0.05	0.04	0.04	0.07	0.06
ROUGE-l	0.02	0.02	0.05	0.04	-0.03	-0.03
coherence-λ=0.0	-0.10	-0.08	0.04	0.04	0.03	0.03
coherence-λ=0.5	-0.16	-0.13	0.06	0.05	0.03	0.03
coherence-λ=0.8	-0.17	-0.14	0.07	0.06	0.00	0.00
coherence-λ=1.0	0.00	0.00	0.07	0.06	0.13	0.11
Ordered BERTScore-v1-n=1	-0.04	-0.03	0.05	0.04	0.06	0.05
Ordered BERTScore-v1-n=2	-0.09	-0.08	-0.07	-0.06	0.00	0.00
Ordered BERTScore-v1-n=3	-0.04	-0.04	0.05	0.04	0.08	0.07
Ordered BERTScore-v1-n=4	-0.05	-0.04	0.04	0.03	0.08	0.06
Ordered BERTScore-v1-n=∞	-0.06	-0.05	0.06	0.05	0.05	0.04
Ordered BERTScore-v2-n=1	**0.04**	**0.04**	0.03	0.02	**0.17**	**0.14**
Ordered BERTScore-v2-n=2	-0.09	-0.08	0.02	0.02	0.11	0.09
Ordered BERTScore-v2-n=3	-0.11	-0.09	0.07	0.06	0.12	0.10
Ordered BERTScore-v2-n=4	-0.11	-0.09	0.05	0.04	0.10	0.08
Ordered BERTScore-v2-n=∞	-0.04	-0.03	0.05	0.04	0.11	0.09

表 2　抽出型要約後の文並び替えの評価

The supreme court on Monday declined to loosen restrictions that prevent individuals convicted of misdemeanor counts of domestic violence from owning a gun.
The case involved two maine men who said their guilty pleas for hitting their partners should not disqualify them from gun ownership. …… The court , in a 6-2 ruling , was unpersuaded by arguments from two maine men who said their past convictions weren't the type that should subject them to a federal ban on firearm possession. ……

図 4　アルゴリズム 1 で要約文書を並び替えた例

	1	2	3	4	5	平均
Zhu らの手法 [27]	0.15	0.20	0.28	0.26	0.14	2.84
Cui らの手法 [5]	0.12	0.18	0.21	0.30	0.21	2.93
ランダム並び替え	0.17	0.23	0.25	0.24	0.13	2.34

表 3　4.1 節の設定での一貫性の人手評価

と比べて評価が大きく下がったため，並び替えを行うことで文書が読みやすくなることがわかった．

6. まとめ

本論文では，抽出型要約における文順序を評価する手法及び，人手で作られた要約文を元に抽出型要約の oracle を並び替える手法を提案した．結果として，既存の評価指標と比べて特に一貫性において人手評価と高い相関が確認できた．今後の課題として，提案並び替え手法の有効性の定量的な検証，他のコーパスや他の並び替え手法での評価指標の検証，並び替えた oracle を教師データとした要約と文並び替えを同時に行うモデルの提案が挙げられる．

参考文献

[1] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, Vol. 34, No. 1, pp. 1–34, 2008.

[2] Jos J. A. van Berkum, Peter Hagoort, and Colin M. Brown. Semantic Integration in Sentences and Discourse: Evidence from the N400. *Journal of Cognitive Neuroscience*, Vol. 11, No. 6, pp. 657–671, 11 1999.

[3] Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management*, Vol. 46, No. 1, pp. 89–109, 2010.

[4] Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4340–4349, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[5] Baiyun Cui, Yingming Li, and Zhongfei Zhang. BERT-enhanced relational sentence ordering network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6310–6320, Online, November 2020. Association for Computational Linguistics.

[6] Peng Cui and Le Hu. Sliding selector network with dynamic memory for extractive summarization of long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5881–5891, Online, June 2021. Association for Computational Linguistics.

[7] Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3739–3748, Brussels, Belgium, October-November

2018. Association for Computational Linguistics.

[8] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics.

[9] Youmna Farag and Helen Yannakoudakis. Multi-task learning for coherence modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 629–639, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of COLING 2012*, pp. 911–926, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[11] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225, 1995.

[12] Yingming Li, Baiyun Cui, and Zhongfei Mark Zhang. Efficient relational sentence ordering network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[13] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1439–1450, Online, July 2020. Association for Computational Linguistics.

[16] Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu. A unified neural coherence model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2262–2272, Hong Kong, China, November 2019. Association for Computational Linguistics.

[17] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 3075–3081. AAAI Press, 2017.

[18] Mir Tafseer Nayeem and Yllias Chali. Extract with order for coherent multi-document summarization. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pp. 51–56, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[19] Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2783–2792, Online, July 2020. Association for Computational Linguistics.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[21] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6209–6219, Online, July 2020. Association for Computational Linguistics.

[22] Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 678–687, Florence, Italy, July 2019. Association for Computational Linguistics.

[23] Yongjing Yin, Fandong Meng, Jinsong Su, Yubin Ge, Lingeng Song, Jie Zhou, and Jiebo Luo. Enhancing pointer network for sentence ordering with pairwise ordering predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 9482–9489, Apr. 2020.

[24] Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. Graph-based neural sentence ordering. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5387–5393. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[25] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[26] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6197–6208, Online, July 2020. Association for Computational Linguistics.

[27] Yutao Zhu, Kun Zhou, Jian-Yun Nie, Shengchao Liu, and Zhicheng Dou. Neural sentence ordering based on constraint graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 14656–14664, 2021.